Frantic race for higher-quality images, at what cost?

Alexandre Schortgen, Lionel Reveret, Guillaume Saulière, Antoine Muller, Thibault Goyallon, Issa Moussa, Jean-François Toussaint

Frantic race for higher-quality images, at what cost?

Application of computer vision models to high level boxing test matches

Schortgen A.[1,2,3]*, Reveret L.[1,2], Saulière G.[3], Muller A.[4], Goyallon T.[1,2], Moussa I.[3], Toussaint J.-F.[3]

[1] Laboratoire Jean Kuntzmann, CNRS UMR 5224, Université Grenoble Alpes, 38400 Saint Martin d'Hères, France
[2] INRIA Grenoble Rhône-Alpes, 38330 Montbonnot-Saint-Martin, France
[3] IRMES - URP7329: Université de Paris Cité, INSEP, Institut de Recherche Médicale et d'Épidémiologie du Sport, F-75012 Paris, France
[4] Univ Lyon, Univ Gustave Eiffel, LBMC UMR_T9406, Lyon, France
*Contact: alexandre.schortgen@inria.fr

Video size

2,4 GB

Context

- Video analysis has become a privileged tool for performance analysis in high level sports. Video recording provides an accurate description of a performance in a non invasive way for a variety of contexts (training or competition).
- Yet, in various disciplines, video analysis is limited to a qualitative visual interpretation. Computer vision can be a tool to provide quantitative analysis by defining objective metrics to describe sport performance.

Problem

A naïve approach to exploiting video data with computer vision models is to consider that the higher the quality (resolution or encoding) of the video, the better the results.
However, implementing systematic video analysis of boxing matches can be challenging, as :

- Training and inference conditions of Deep Learning models may differ.
- Video quality may be low due to sourcing or transfer encoding.
- Storage capacity may limit the volume/quantity of matches analysed.

How do Computer Vision algorithms perform on lower quality images?
To what extent can we optimize video storage without compromising analysis accuracy?

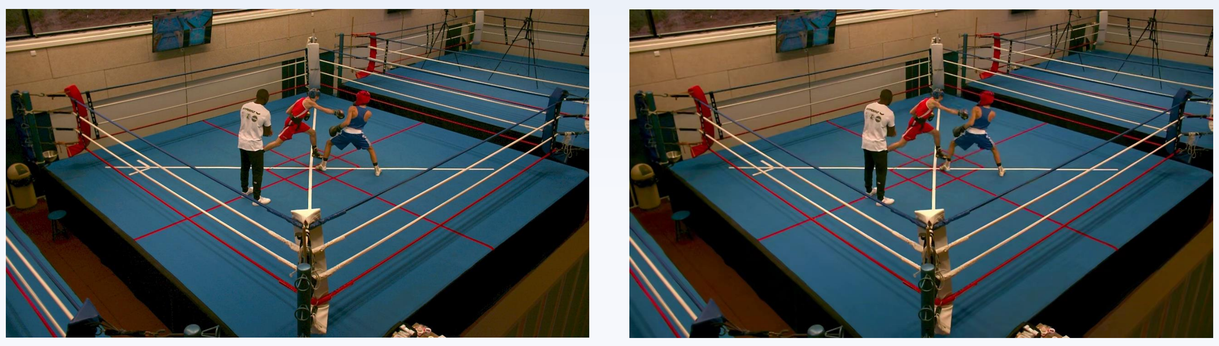

Fig. 1. Comparison between 2 frames of a given moment of the match – Left: Reference video – Right : Compressed video encoded at 500kB/s.

Method

- Video data : 1 round of 3min (test match of elite boxers)
- Monocular video analysis
- Compression bitrates investigated : 500, 1500, 3000, 6000, 10.000 and 100.000 kB/s.
- The following OpenPose[1] body_25 outputs were computed on the same monocular video for all the previous bitrates:
  - Number of poses reconstituted per frame
  - Position of joint centres
  - Confidence score on joint detection
- Average number of poses per video : 34.075 ± 228
- Average number of joints detected per video : 633.258 ± 7.395
- Reference video characteristics :
  - resolution : 1270x720 px
  - frame rate : 60 fps
  - number of frames : 11.291
  - bitrate : 1,9 MB/s
  - Size : 4,5 G
- Joint location error between compressed and reference videos for a joint j at frame t :

$$\left\| p_{t,j}^{ref} - p_{t,j}^{comp} \right\|_2 = \sqrt{(x_{t,j}^{ref} - x_{t,j}^{comp})^2 + (y_{t,j}^{ref} - y_{t,j}^{comp})^2}$$

- Statistical analysis:
  - Pearson's χ^2 goodness-of-fit test was used to compare the distributions of the number of poses detected per frame between reference and compressed videos.
  - Wilcoxon test was used to compare the medians of the confidence score distributions.

Results

1 Number of poses per frame

The distributions of per-frame pose numbers obtained on 11261 frames for downgraded videos are significantly different from the reference distribution and are also all different from one another.

To better quantify this difference, we investigated the proportion of frames for which the pose number does not match the reference results.

Encoding (kB/s)	500	1500	3000	6000	10 000	100 000
Frames w/ ≠ nb of poses (%)	15.8*	14.5*	13.7*	13.3*	13.4*	13.1*

Table 1. proportion of frames for which the number of poses detected differs from the reference.
** : distribution significantly different from the reference (p-value < 0.01 after Bonferroni correction)*

Decreasing the video quality modifies the output of the model.

Dividing the video size by a factor 200 only increases the incorrect amount of poses by 3%.

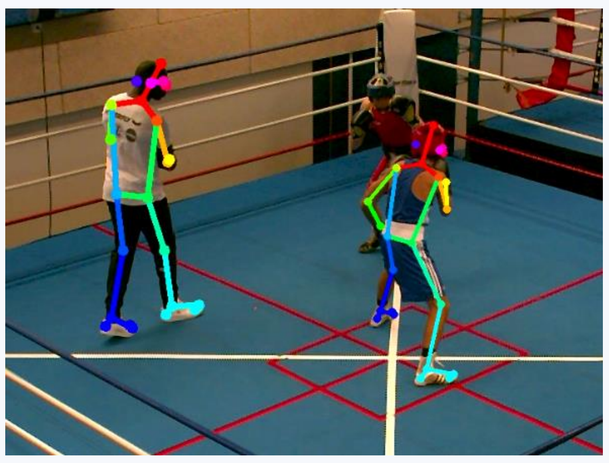

Fig. 2. Example of erroneous pose estimation due to partial occlusion of the red boxer.

2 Position of joint centres

For most of the bitrates investigated, 95% of the joint centres are located within a 6-px distance from the reference joint positions.

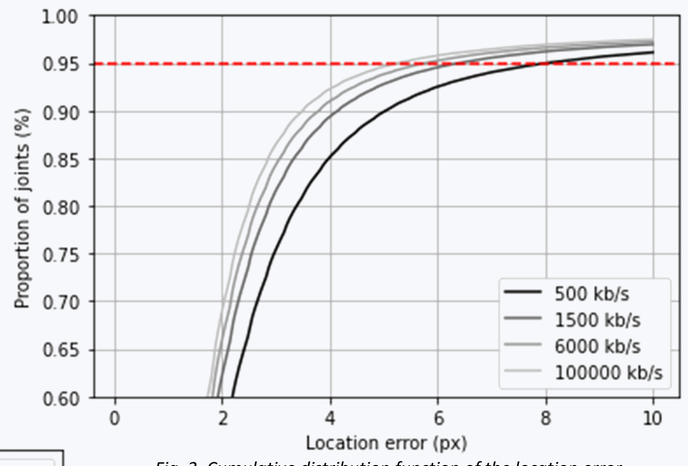

Fig. 3. Cumulative distribution function of the location error.

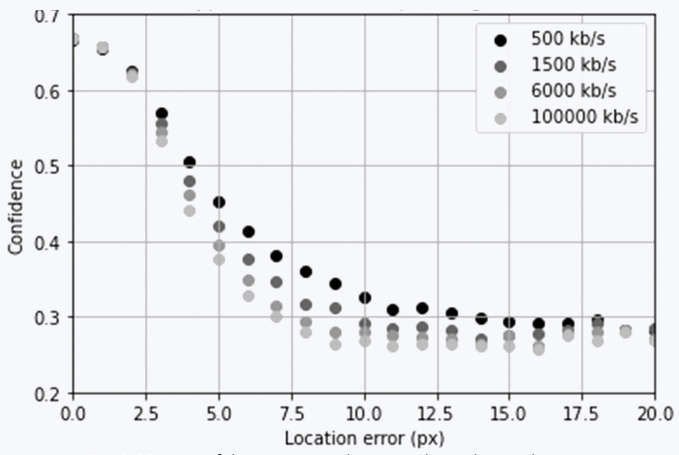

Fig. 4. Mean confidence on joint detection depending on location error (rounded).

A Spearman rank-order correlation coefficient of -0.34 (p-value<0.01) highlights a significant decreasing monotonic correlation between the location error and the confidence score of joint detection.

When analysing the results by joint type, higher errors appear to be located at the less visible parts of the body (hip and thighs).
Moreover, the median confidence score on all the joint detections decreases significantly with the compression (from 0.71 to 0.66).

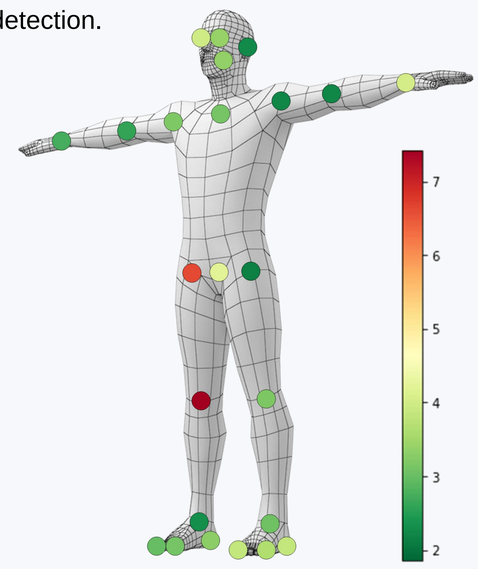

Fig. 5. Mean location error (px) grouped by joint.

Location error < 6 px for 95% of the data.

Higher errors occur on uncertain, thus less important, detections.

237 MB

142 MB

71 MB

36 MB

12 MB

0

Main challenges for accurate pose estimation
- Boxing outfit & movement speed
- Close-range fighting and clinch phases
- Monocular 2D data

Is a 6-px error on joint location acceptable ?
- × Markerless motion capture
- ✓ Overall localisation of the boxers inside the ring

Acknowledgements : The authors would like to thank the boxers who participated in the experiment and the French Boxing Federation for their collaboration.

References : [1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, et Y. Sheikh, IEEE, « OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields », 2019.